DOCUMENT RESUME

ED 260 116 TM 850 440

AUTHOR	Holmes, Susan E.
TITLE	Test Equating Issues in Certification and Licensure Testing.
PUB DATE	1 Apr 85
NOTE	22p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Speeches/Conference Papers (150) -- Viewpoints (120)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Certification; Competence; Criterion Referenced Tests; Cutting Scores; Difficulty Level; *Equated Scores; Knowledge Level; *Occupational Tests; Scaling; *State Licensing Boards; Test Construction; Testing; *Test Selection
IDENTIFIERS	Open Testing; Test Disclosure; *Test Equivalence; Unidimensionality (Tests)

ABSTRACT

Open testing and test disclosure legislation has fostered a renewed interest in the problems and issues associated with test equating. This paper focuses on a specific application of test equating, that of professional and occupational testing for certification and licensure. Test equating is the process of making test scores from different forms of the same test comparable. The role of test equating in the credentialing examination process, its usefulness for maintaining consistent standards, and monitoring performance trends across time are discussed. The issues which must be considered when implementing test equating in a credentialing examination program are described. Issues are considered as test development or technical concerns. This paper is designed to provide information for credentialing board members who hold the major responsibility for preparing, administering, and scoring credentialing examinations, as well as non-technical persons involved in the credentialing process. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED260116

Test Equating Issues in Certification and Licensure Testing

Susan E. Holmes, Ph.D.

The Psychological Corporation

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Holmes, Susan E.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A Paper presented in a symposium entitled
"Issues in Certification and Licensure"
presented at
The Annual Meeting of
The American Educational Research Association
held in Chicago
April 1, 1985

BEST COPY

TM 850 440

Abstract

The advent of open testing and test disclosure legistation has fostered a renewed interest in the problems and issues associated with test equating. This paper focuses on a specific application of test equating, namely that of professional and occupational testing for certification and licensure. First, the role of test equating in the credentialing examination process is considered. Its usefulness for maintaining consistent standards and monitioring performance trends across time are discussed. Second, the issues that must be considered when implementing test equating in a credentialing examination program are described. Issues are considered as either test development concerns or as technical concerns. The audience for the paper is credentialing board members who hold the major responsibility for preparing, administering and scoring credentialing examinations and other non-technical persons involved in the credentialing process.

BEST COPY

Test Equating Issues in Certification and Licensure

Introduction

Until recently, test equating was viewed as a somewhat obscure and specialized set of statistical methods of interest only to testing organizations and selected measurement specialists. This situation has changed, however, with the advent of open testing and test disclosure legislation. In recent years, test equating has been thrust into the public arena, coming under the close scrutiny of educators, legislators, and the general public.

Beginning as early as the 1950's, contributions to the research literature on test equating can be found (Lord, 1950, Flanagan, 1951; Lord, 1955). Over the years, numerous articles have been written that describe test equating methodology or report on the results of comparisons among a variety of test equating techniques. Angoff (1971) provides an excellent summary of the methodological literature to that time. A more current review can be found in a book entitled Test Equating (Holland & Rubin, 1982).

One application of test equating that has received little attention is in the area of professional and occupational testing for certification and licensure. This is not surprising because recognition of the special problems associated with testing in this area has emerged only recently. In the final report on Standards for Educational and Psychological Testing, professional and occupational testing is considered in a separate chapter for the first time (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, November 1984).

The purpose of this paper, and the symposium of which it is a part, is to highlight issues in professional and occupational testing for certification and licensure. This paper will focus on test equating. Two topics will be considered here. First, the role of test equating in credentialing examinations will be examined. Second, the issues that must be addressed in selecting and implementing test equating in a credentialing examination program will be considered.

Many technical presentations about test equating methodology can be found elsewhere in the research literature. The audience for this paper, therefore, is individuals who hold the major responsibility for preparing, administering, and scoring credentialing examinations, namely certification and licensing board members and other non-technical persons involved in the credentialing process. The paper is intended to assist these individuals in working through some of the problems and technical concerns that must be considered before selecting and implementing a test equating technique.

BEST COPY

Why Credentialing Examinations Warrant Special Attention

As mentioned above, some recognition of the expanding role of professional and occupational testing for credentialing has recently emerged. This area of testing warrants special treatment for a number of reasons, some of which will be enumerated below.

First, the outcomes of credentialing examinations have an enormous impact on the lives of individuals, on institutions, and on society in general. Such tests have associated with them the potential threat of serious harm to the public as well as to the individuals who have invested a considerable amount of time, money and effort preparing for a particular profession or occupation. It therefore seems particularly critical that the testing practices associated with credentialing examinations be impeccable and that they represent state-of-the-art technology.

Second, the number of professions and occupations currently requiring some kind of credentialing is greater than ever before, and the figure continues to increase. In 1982, the number of occupations regulated by states was estimated to be over 800 (Shimberg, 1982). This figure does not include professions or occupations that are subject to federal or local regulation. Of more interest here is the fact that the increase in credentialing has been accompanied by an increase in the amount of testing being conducted for this purpose. This situation creates a corresponding increase in the number of opportunities for serious harm as well as legal entanglement.

Third, the quality of credentialing examinations in the past has not always been high (Shimberg, Esser & Kruger, 1972). This situation is, in many ways, a result of special circumstances associated with this type of testing. For example, credentialing board members usually hold the major responsibility for these examinations. They are often selected for this role because they are considered to be qualified practitioners in their particular field. Such individuals usually do not also possess training and experience in test construction or the statistical analysis of test items. Further, many credentialing programs involve small numbers of applicants, eliminating the benefits of economy of scale that might defray the high costs associated with operating a sound testing program.

This is certainly <u>not</u> to imply that all credentialing examinations are poor quality. A great many boards have developed a considerable amount of expertise in test development, often making use of testing consultants or staff from centralized state testing units. In other instances, boards have formed consortia to lower testing costs or have subscribed to national testing programs.

However, the issue of quality must be considered because it has important implications. Boards, the public and the legislature often assume that credentialing examinations are <u>absolutely valid</u> measures of competence. As a result, even though an applicant may satisfy all other requirements prior to sitting for a credentialing examination, he or she will not be given the opportunity to combine their test scores with this other information to modify

the decision rules in particular cases should the individual fail to achieve the passing score by as little as one raw score point.

Finally, professional and occupational testing warrants special attention because its purposes and goals are quite different from those in other areas of testing. Consider for example, employment testing, an area in which professional and occupational testing has been included in the past. The purpose of employment testing is to make predictions or decisions about individual personnel actions. Its focus, therefore, is on maximizing productivity and the individual's aptitude to learn a specific job. The focus in professional and occupational testing, however, is on the level of knowledge and the skills necessary to reasonably assure the public that individuals are competent to practice.

Other differences exist between professional and occupational testing and personnel testing. Individuals taking credentialing examinations have usually completed their training and are seeking to be deemed qualified to practice in a broad field. Further, there are no explicit limits on the number of individuals who can be considered qualified. Individuals taking employment or selection examinations, on the other hand, are competing for a limited number of available openings for a specific job. In the latter situation, the required level of aptitude is clearly influenced by supply and demand.

Professional and occupational testing has still other distinguishing characteristics which have specific implications for test equating. First, most credentialing examination programs require multiple test administrations each year. In order to ensure fairness, a new version of the test is needed for each administration, and strict security measures must be enforced. Second, these examinations tend to be quite lengthy, sometimes involving one or more days of testing for each administration. Finally, the timely return of testing results is a necessity. In combination, these requirements make many of the traditional test equating paradigms extremely difficult, if not impossible, to implement. A more detailed discussion of these issues and their impact on test equating will be given later in this paper.

BEST COPY

Some Relevant Background

Licensing is the process by which an agency of the government grants permission to an individual to engage in a given occupation. Its primary purpose is to protect the public from incompetent practitioners. It accomplishes this end by imposing requirements the purpose of which is to reasonably ensure that those licensed possess knowledge and skills in sufficient degree to perform important occupational activities safely and effectively.

Before a license is granted, the applicant, must satisfy certain requirements set forth by law. Included among these are such things as training and experience, minimum age, academic degrees or years of formal education, and a period of residency within a particular state.

The applicant for licensure must also demonstrate that he or she is competent enough to ensure that the public health, safety and welfare are reasonably well protected. In many cases, the minimum level of competence required is defined in terms of a test score on a licensure test, and as a rule, only applicants who satisfy the stated prerequisites are allowed to sit for the test. One of the most difficult tasks facing credentialing boards today is establishing the minimum level of competence.

The responsibility for examining applicants usually rests with the licensing board. In some cases, boards may actually assume responsibility for preparing, administering and scoring the examinations. In such situations, boards may develop their own testing capabilities by employing professional staff with special expertise in measurement. In other instances, they may rely on national testing organizations to assist in the development of their examinations. In either case, the examination process is frequently the responsibility of boards comprised largely of members of a particular profession.

In contrast, certification is the process by which a governmental or nongovernmental agency or association grants authority to an individual to use a specified title. Like licensure, the primary purpose of certification is to protect the public. Certification is a means of granting recognition to individuals who have attained some required degree of knowledge and skill in a given field. Unlike licensure, however, certification does not prohibit individuals from engaging in an occupation. Further, certification requirements are usually set at a level well above the minimum necessary to protect the public from incompetent practitoners. Some exceptions to this occur in instances where certification is set at the entry level.

The range of occupations currently covered by some form of certification or licensing is wide. It includes many of the traditional professions such as medicine, law and teaching, as well as technical and specialized occupations such as air traffic control, computer programming, cosmetology, mechanics and real estate. This paper will deal primarily with licensed occupations, although many of the issues and concerns described apply equally to alternative modes of credentialing such as certification.

The Role of Test Equating In Credentialing Examinations

One of the most difficult and important tasks facing individuals responsible for credentialing is defining the level of competence required for certification or licensure. This important activity, referred to as standard-setting, has received a considerable amount of attention in the literature. A large body of research has been published describing this methodology, delineating the issues involved in determining standards and translating them into passing scores on tests, and contrasting the results produced by different techniques. An entire issue of the Journal of Educational Measurement was devoted to the subject (1978) and a number of extensive reviews have been published (Hambleton, Powell & Eignor, 1979; Sheppard, 1979, 1981).

Much of the standard-setting literature has focused on criterion-referenced tests and the use of standard-setting techniques such as those proposed by Angoff (1971), Ebel (1972), and Nedelsky (1954). These techniques consist of a variety of procedures for arriving at a performance standard for a test, usually a percent correct score on a test. Such approaches are referred to as "criterion-referenced" because they are based on absolute, rather than relative, judgments about test content or individuals. As such, they are heavily dependent on the specific items included in the test being considered.

While much has been written about techniques for establishing standards for minimally acceptable levels of competence, very little has been published about how to ensure that these standards, once established, are maintained over time. Most of the techniques yield a passing score that is appropriate only for the examination to which the technique was applied. These techniques must therefore be applied to all subsequent examinations for which passing scores are desired.

This situation presents some serious problems in professional and occupational testing. Credentialing examinations are frequently administered several times a year. Ideally, a new version, or form, of the test is administered each time. Each form is similar in content and format, in the number and type of questions it includes, and in the range of difficulty of the questions since each is developed from the same set of test specifications. In order to maintain fairness and security, however, each new form also contains different test questions. This practice prevents individuals who take the examination at a later date from having an obvious and unfair advantage over those who took it earlier. It also contributes to unintended differences in overall test difficulty.

Because each new form of the test contains a different set of test questions, a passing score must be determined for each in such a way that all passing scores represent the same level of competence. Conducting a criterion-referenced standard-setting study each time a new form is administered, however, can be prohibitively expensive and time consuming. Even more disconcerting is the fact that quite different results have been reported in studies where the same technique was applied to the same examination more than once (Andrew & Hecht, 1976; Halpin, Sigmon & Halpin, 1983; Mills & Barr, 1983; Saunders, Ryan & Huynh, 1981). These findings also raise questions about the consistency of passing

scores obtained from the application of the same technique to different forms of a test.

An alternative approach to the standard-setting problem is the use of normative information to select passing scores. Although much has also been written about this more traditional approach, it is intrinsically unappealing in professional and occupational testing for several reasons. First, the critical issue in credentialing is whether the individual is a safe and effective practitioner and not whether he or she is safe and effective compared to the other applicants who happen to be sitting for the examination at this administration. Competence has an absolute, rather than relative, nature. Further, a norm-referenced approach does not take into account differences among populations of applicants sitting for different administrations or differences in test difficulty. The level of competence of applicants in the lowest 25 percent at one administration, for example, may represent a quite different standard when applied to applicants sitting for the examination at a different time of the year. Equating successive forms of an examination offers a partial solution to the standards problem.

Test equating is the process of making test scores from different forms of the same test comparable. Stated more formally, test equating is the process of converting the system of units of one test to the system of units of another so that the information obtained from the two, after conversion, will be equivalent (Angoff, 1971).

Although content and construct interpretations of a test are of primary importance, test equating contributes to the meaning of test scores. It is necessary to equate different forms of the same test because unintended differences in overall test difficulty from form to form can give an unfair advantage to those individuals who happened to take the easier form. The conversion of scores may be accomplished either by applying an equation that relates the scores on one form to the scores on the other or by looking up these values in a table developed using the equation.

One role for test equating in professional and occupational testing, then, is to ensure that consistent standards of competence are applied across all administrations of an examination. The particular form of the test taken should be a matter of indifference to the individuals taking the test and to those using the test results (Lord, 1980). This will not be the case if unequated raw scores are used to evaluate competence. Individuals taking the easier test will have an unfair advantage without test equating. Because the purpose of licensure is to protect the public, this situation is unacceptable.

Test equating represents only a partial solution to the problems of standard-setting for several reasons. First, its value is based on the assumption that an appropriate passing score has already been determined. The current state of the art with respect to standard-setting does not necessarily support this assumption. Further, the existing equating methodology has associated with it certain requirements and assumptions that make its application extremely difficult in the context of credentialing examination programs. These concerns will be explored in more detail later in this paper.

A second, and equally important, role for test equating in professional and occupational testing is to monitor performance trends across time. If successive forms of the examination are equated, changes in the number or percent of applicants awarded credentials will reflect real changes in the competence of the individuals sitting for the examination and not differences in the difficulty of the test itself or the standard being applied. Such information is extremely useful for planning, curriculum evaluation and policy decision-making.

Test Development Issues in Test Equating

Test equating has many facets. It is most often viewed as the process of making statistical adjustments to the scores obtained on different forms of the same test to compensate for differences in relative difficulty. This is, however, only one aspect of the process. Test equating also consists of applying a number of test construction techniques throughout the test development process. Both aspects of the test equating process are equally important and both will be considered in this paper, beginning with the test development component.

The issues presented here and in the next section of the paper have been selected because of their potential impact on the accuracy of test equating results. Their inclusion in one or the other of the two sections that follow is in some instances a matter of preference . The issues cover a wide range of topics. In some cases, they may be rather obvious or already well established. These issues are included for the sake of completeness. In other cases, they may be more speculative in nature, intended to suggest possible approaches rather than to provide conclusions. In either event, the comments are based on a review of the test equating literature and several years of experience working as a psychometric consultant responsible for designing and implementing test equating programs for credentialing boards.

1. Do the tests being equated measure the same thing?

This issue involves perhaps the most basic requirement implicit in the definition of test equating. It is theoretically impossible to equate tests that measure different things. Test equating makes sense only if the forms of the test being equated measure the same trait or attribute. This issue is considered important enough to be included in the new Standards for Educational and Psychological Testing (1984). Reporting data on the parallelism of tests being equated is a standard that should be met by all tests before their operational use.

This concern can be minimized by using well-defined test specifications to guide the test development process and thoroughly documenting the process. It becomes a problem, however, when such specifications either do not exist or when they must be changed over time. The latter situation arises when, for example, test specifications must be modified to reflect current practice. In such situations, it may be necessary to make the scores on such tests comparable in some particular sense for a particular population. However, care should be taken to avoid the implication that such tests have been "equated", since by definition this is possible only with tests measuring the same thing. A more appropriate description is that the scores have been "scaled" to achieve comparability. The accuracy of the equating results obtained in such situations will depend on the degree to which the two tests differ with respect to what they measure.

2. Are the tests unidimensional?

Not only must the two tests being equated measure the same thing, they must also measure only one thing. Tests that measure only one trait or attribute are referred to as "unidimensional".

Unidimensionality is a basic prerequisite for test equating. This assumption may not be tenable, however, for some credentialing examinations. Credentialing examinations often cover a wide range of knowledge and skills even though performance may be summarized with a single test score. In such situations, the tests may need to be subdivided into smaller more homogeneous subtests, each summarized by a single score. Test equating would then be performed at the subtest level.

The issue of unidimensionality is considered here, rather than in the section on technical issues in test equating, because it is basically a test development issue. No completely satisfactory statistical technique for evaluating the unidimensionality of multiple choice tests currently exists. However, factor analysis, tetrachoric correlations, phi correlations, and the Kuder-Richardson Formula 20 have all been used to obtain evidence of the tenability of this assumption (Hambleton & Swaminathan, 1985).

3. Are the difficulty levels of the tests similar?

More accurate equating results will be obtained from tests that do not differ greatly in difficulty. The technical problems associated with vertical equating are well documented. However, vertical equating situations are not frequent in professional and occupational testing. Here, test equating is usually necessary because of unintended differences in difficulty. A good rule to keep in mind is that it is generally easier to construct tests carefully than it is to try to compensate for the results of poor test construction at the time the tests are being equated.

Equating accuracy will also be affected by the distribution of the item difficulties on the two tests. Tests having similar distributions yield more accurate results than tests having very different distributions.

Another related issue is the appropriateness of the difficulty levels of the tests being equated. Generally speaking, tests that are too difficult yield unsatisfactory equating results, particularly at the lower end of the score distribution. This situation can be a serious concern for licensing examinations since important decisions are often made at the lower end of the score distribution. Similarly, tests that are too easy yield unsatisfactory equating results at the upper end of the score distribution. This situation is more of a concern in certification where decisions tend to be made at the upper end of the score distribution.

Finally, equating accuracy is affected by the number of distinct scores possible on the tests being equated. While this may be considered to be a scoring issue, it is also related to the difficulty of the tests. If only a small number of distinct

BEST COPY 12

scores are possible, there will be heavy bunching of the test scores, making equating more difficult. Potthoff (1982) suggests several ways to avoid this problem. Included among these is the use of formula scoring, item response theory scoring, and differential item weighting of both correct and incorrect answers.

4. Are the tests equally reliable?

Just as it is impossible to equate scores on tests that measure different things, test forms that differ in reliability cannot be equated in the strict sense. If this were not the case, there would be no reason to build reliable tests. Unreliable tests could be made to yield reliable scores simply by equating them to reliable tests.

More importantly, however, two unequally reliable tests can not be equated because the result will not satisfy the definition of equating given earlier. In this situation, which test is taken is no longer "a matter of indifference" to individuals taking the test or to those interpreting their scores. Incompetent individuals have a better chance of achieving the passing score if they take the less reliable test if they obtain a favorable measurement error. On the other hand, competent individuals would prefer the more accurate test, thus minimizing the chance of obtaining a score below the passing score due to an unfavorable measurement error (Lord, 1977).

5. Will the format of the items impact the test equating results?

An item type frequently encountered in credentialing examinations is the "case" format. Here, one or more test items are linked to a common passage. Although the face validity of such items is apparent, their use greatly reduces flexibility and makes it more difficult to meet test specifications. It has been suggested that such item types may indirectly cause equating problems and that it would be wise to avoid their use wherever possible (Potthoff, 1982).

6. Are all nonoperational sections of the tests indistinguishable?

Some credentialing examinations include one or more nonoperational test sections. Such sections might contain anchor items to be used in the equating or experimental items being field tested for some future form of the test. An effort should be made to conceal the identity of such sections from examinees since motivation plays an extremely important role in the quality of the item information obtained during credentialing examinations.

7. Are the time limits for the tests appropriate?

Inappropriate time limits can have an adverse impact on the accuracy of equating results. Allowing applicants insufficient time to complete the test or administering tests that are too lengthy can result in large numbers of omitted items or poor quality item information due to fatique. Both situations can cause equating problems, particularly if an item response theory model is used to perform the equating.

The second aspect of the test equating process involves the selection and implementation of a test equating technique. An test equating technique is an empirical procedure for determining a statistical adjustment to be applied to the scores on one of two forms of a test to compensate for unintended differences in difficulty. As such, it involves a design for data collection and a rule for determining the adjustment to be applied.

Generally speaking, there are three basic designs for data collection and three rules for determining the adjustment in test equating. Any of the three data collection designs can be used with any of the three adjustment rules. For the sake of completeness, a brief description of the three data collection designs and the three rules for determining adjustments is given next. For a more detailed and technical presentation, see Lord (1975, 1980).

The three designs for data collection are referred to as the single-group design, the equivalent group design, and the anchor-test design. They can be described as follows:

> Single-group design: The two tests are both administered to the same group of applicants, the order of presentation being counterbalanced;
>
> Equivalent-group design: Each test is administered to one group of applicants. The applicants in each group are identified from a larger group by random sampling or some other method so that the groups are approximately equivalent.
>
> Anchor-test design: Each test is administered to one group of applicants, and a common anchor test is administered to both groups. The anchor test may be a subset of items contained in both tests or it may be a separately timed test that each applicant takes in addition to taking one of the two tests being equated.

There are three widely used techniques for determining the statistical adjustment to be made in test equating. They can be described as follows:

> Equipercentile equating: Adjust the scores from the two tests so that they correspond to the percentile ranks in some reference group;
>
> Linear equating: Adjust the scores from the two tests so that they correspond to standard deviations from the mean in some reference group;
>
> Item response theory equating: Adjust the scores from the two tests so that they correspond to the levels of the latent trait (competence) measured by each test.

Some of the more technical issues related to selecting and implementing test equating follow next. Again, their inclusion in this section may be more a matter of preference than a reflection of reality.

BEST COPY

1. Which data collection design should be employed?

The most appropriate data collection design for a given test equating situation will, of course, depend on the specifics of that particular application. Such a selection would be best made in conjunction with a measurement specialist. However, a few comments will be offered regarding the general usefulness of the three basic designs described above in professional and occupational testing for certification and licensure.

Some of the special needs associated with credentialing testing were described earlier in this paper. Included among these were the need for multiple new test forms each year, tight test security to ensure fairness, and timely reporting of test results. When coupled with the fact that credentialing examinations tend to be quite lengthy, these circumstances make the implementation of the first two data collection designs impractical.

The problems associated with the single-group design are obvious. Administering two different forms of a credentialing examination to one group of applicants, randomly assigning the order of administration, is not practical in most instances because of time considerations. Further, this approach has legal implications associated with it. It could be argued that having some applicants take their credentialing examination first while others take it second is unfair, either because of the advantage of more practice or the disadvantage of fatigue. While such an administration might be conducted prior to the actual testing using a sample of applicants, security concerns and the need to conduct such a study several times each year would seem to preclude this as a solution.

The second data collection design, the equivalent-group design, has similar problems associated with its use in credentialing testing. As with the first design, security concerns make its implementation prior to the actual test administration undesirable. Further, the need to conduct such studies several times a year is problematic. More importantly, however, are the legal implications associated with the use of this design during the actual test administration since it requires that randomly selected groups of applicants take different tests, one of which was used in some previous administration. One advantage of this design is that it requires each applicant take only one test.

The anchor-test design is the most promising for credentialing testing. There are several reasons for this. First, this design requires each applicant to take only one test. Second, the data necessary to perform the equating can more easily be collected during the actual test administration. Third, the issue of security, while still present, is much less of a concern, since only a subset of items in the examination, the anchor set, has appeared in a previous administration. Since performance on the anchor test is used to reduce equating bias resulting from differences in the level of ability of the two groups participating in the equating, the effectiveness of this design depends on the correlation between the anchor test and the two tests being equated.

Regardless of which data collection design is selected, the details of that design should be well documented. The information provided should include things like the sampling procedures employed, the sample size, and the characteristics of the

sample. If an anchor test design is selected, the characteristics of the anchor test should be described and its contents summarized and compared to the tests being equated.

2. Which adjustment rule should be applied?

Once an appropriate data collection design has been identified, a rule must be selected for determining the statistical adjustment to be applied. This decision is best made in conjunction with a measurement specialist familar with the technical qualities of the tests being equated. However, the fact that most credentialing examinations involve only "horizontal" equating makes the decision somewhat easier. In horizontal equating, adjustments are made to the scores on two tests in order to compensate for unintended differences in their difficulty or for differences in the groups of applicants (Slinde & Linn, 1977). In "vertical" equating, the two tests being equated intentionally differ in difficulty. The selection of an adjustment rule is easier in horizontal equating situations because all three methods have been found to work reasonably well in this situation.

Much research has been conducted to date in which the results obtained from the application of one or more of the three methods in horizontal equating situations are compared and contrasted. (See, for example, Lord, 1975, 1977 or Petersen, Marco & Stewart, 1982.) No single method has been found to be consistently superior to the others. In fact, the findings suggest that if the tests are reasonably reliable and of comparable difficulty, and if the groups taking the tests are reasonably equivalent, then any of these methods will yield satisfactory results (Skaggs & Lissitz, 1981).

While all three equating methods have been found to yield quite similar results in horizontal equating situations, the third method based on item response theory (IRT) offers other advantages that make it particularly attractive. These stem from the fact that once a set of test questions have been fitted to an IRT model, it becomes possible to deal more effectively with several other measurement problems. Because the focus of IRT is at the item rather than test level, this measurement approach makes it possible to construct credentialing examinations that maximize the accuracy of decisions at critical points along the competence continuum. Another advantage that IRT affords is a separate estimate of measurement error for each level of competence, rather than only one estimate which must then be applied to all levels. This feature makes it possible to determine how precise an estimate of competence is, say at the passing score, and more importantly, to control the size of the measurement error at such critical points along the continuum during the test construction process.

As one might expect, there are several issues that must be resolved before the benefits of IRT can be realized. In some credentialing examination programs, these may make the application of IRT impossible. For example, fairly large samples of applicants are needed in order to obtain stable item parameter estimates. Some credentialing programs simply do not have a broad enough applicant base to make the implementation IRT feasible. This includes both inadequate numbers of applicants as well as insufficient resources in general. For a more detailed presentation on IRT and the issues associated with its

BEST COPY

implementation, see the special issue of the Journal of Educational Measurement (Summer, 1977), or one of the several textbooks currently available on the subject (Hambleton, 1983; Hambleton & Swaminathan, 1985; Lord, 1980; Warm, 1978; Wright & Stone, 1979).

IRT is being successfully employed in a number of credentialing examination programs at the present time. For example, the National Council Licensing Examination (NCLEX) for registered nurses and practical nurses, sponsored by the National Council of State Boards of Nursing, uses IRT both for test equating and test development. These two licensing examination programs test over 200,000 nurses each year across the nation.

Regardless of which general adjustment rule is used, the need to document must again be addressed. The particulars of the equating method employed should be specified as well as the rational for its selection. Any statistical procedures used during the equating, like smoothing techniques or adjustments for extreme scores, should also be described along with a rationale for their use.

3. How will the adequacy of the equating be assessed?

The accuracy of the equating results obtained from the particular equating method being used should be evaluated periodically. One source of information would be a comparison of the results obtained from the application of a second method. A visual inspection of equating results across the entire continuum can help identify systematic differences. A comparison of summary statistics across tests is another valuable source of information.

An important consideration in evaluating the adequacy of a test equating is the similarity of the two tests. Form-to-form differences in the number of items in a category and in the distributions of item statistics can be very informative.

4. What are the potential sources of error in test equating?

Little has been said yet about the issue of equating errors. However, whenever test equating is performed, the resulting conversion will have errors associated with it. These may be attributed to either the data collection design employed, the adjustment rule applied, or both. Potential sources of error should be identified and then estimated at various points along the continuum. In some cases, procedures for estimating the magnitude of equating errors are routine. In others, special analytical techniques will be required.

5. What score scale is most appropriate for score reporting?

Raw scores on a test are expressed in units that are the result of arbitrary features of a particular test. It is therefore often convenient to report raw scores in terms of a derived score scale to enhance their meaning. This conversion process is called scaling.

Reporting scores from different administrations of an examination on the same derived score scale has several advantages. First, it allows a consistent value to be assigned to important scores like the passing score. Second, it avoids the necessity of lengthy explanations about why different scores are required to pass different examinations.

Although scales can enhance the interprebility of test scores, they can also lead to confusion and misinterpretation. For this reason, it is recommended that care be taken in selecting a scale. The scale selected should be consistent with the intended purposes of the test and described in detail (Standards for Educational and Psychological Testing, 1984).

BEST COPY

Conclusions and Future Challenges

The advent of open testing and test disclosure legislation has fostered a renewed interest in the problems and issues associated with test equating. The focus of this paper was on a specific application of test equating, namely in the area of professional and occupational testing for certification and licensure. Such an application warrants special attention because of the enormous impact credentialing examinations have on the lives of individuals, on institutions, and on society in general.

The paper addressed two topics. The first was the role of test equating in the credentialing process. Two distinct functions were identified. One was to ensure that consistent standards of competence are maintained across time. A second, and equally important, role was to monitor performance trends across time. Such information is essential for planning, curriculum evaluation and policy-decision making.

The second topic considered in the paper was issues that must be addressed in selecting and implementing test equating in a credentialing examination program. Since test equating has two major components, the issues were presented as either test development concerns or as technical concerns. In either case, the issues identified were included because of their potential impact on the accuracy of test equating results.

Many of the test development issues cited dealt with the need to employ sound test construction practices during the test development process. The basic theme was that it is easier to construct a good test than it is to try to compensate for a poor one at the time that equating must be performed. "Good" tests here would be ones that measure exactly the same thing, are reasonably reliable, and of comparable difficulty.

Some of the technical issues presented were less easily resolved. For example, several of the data collection strategies traditionally employed in test equating were found to be impractical for credentialing examinations because of the special requirements of such testing programs. Other issues, like the one invclving unidimensionality, lack good technical solutions at the current time. In an ideal psychometric world, the tests to be equated would be similar in all important respects and random samples of applicants would be readily available for equating purposes. Unfortunately, this is not the case in the real world in general, and in professional and occupational testing for certification in particular, where program policy and legislative requirements rule the day.

IRT was discussed as a particularly promising approach for credentialing examination programs. The advantages that it offers, however, may be out of reach for many programs because of limited numbers of candidates and inadequate resources in general.

The obvious conclusion is that test equating has a signifcant role to play in professional and occupational testing for certification and licensure. The discussions presented indicate, however, that certification and licensing boards will need more technical support and resources if they are to use test equating procedures routinely. The bigger issue to consider is whether the cost of implementing such procedures exceeds their utility in instances where boards test few applicants each year or have limited resources. Perhaps other ways of determining competence should be explored in these cases, ways that place less emphasis on test scores but rather rely on other indicators of competence.

Test equating remains a fertile area for future research. The current renewal of interest may provide solutions to some of the problems now faced in professional and occupational testing for certification and licensure. It is hoped that this paper, and the symposium of which it is a part, has made some contribution toward that end.

BEST COPY

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. Standards for educational and psychological testing (Final Report of the Joint Technical Standards Revision Committee). Washington, DC: American Psychological Association, November 1, 1984.

Andrew, B.J., & Hecht, J.T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W.H. Scales, norms and equivalent scores. In R.Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Flanagan, J.C. Units, scores and norms. In E.F. Lindquist (Ed.), Educational Measurement. Washington, DC: American Council on Education, 1951, 695-763.

Halpin, G., Sigmon, G., & Halpin G. Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 1983, 43, 185-196.

Hambleton, R.K. (Ed.). Application of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 1983.

Hambleton, R.K., Powell, S., & Eignor, D.R. Issues and methods for standard setting. In R.K. Hambleton & D.R. Eignor (Eds.), A practitioner's guide to criterion-referenced test development, validation, and test score usage. Amherst, MA: Laboratory of Psychometric and Evaluative Research, University of Massachusetts, 1979.

Hambleton, R.K., & Swaminathan, H. Item response theory: Principles and applications. Boston, MA: Kluwer Nijhoff Publishing, 1985.

Holland, P.W., & Rubin, D.B. Test equating. New York, NY: Academic Press, 1982.

Lord, F.M. Notes on comparable scales for test scores (ETS RB 50-48). Princeton, NJ: Educational Testing Service, 1950.

Lord, F.M. Equating test scores: A maximum likelihood solution. Psychometrika, 1955, 20, 193-200.

Lord, F.M. A survey of equating methods based on item characteristic curve theory (RB-75-13). Princeton, NJ: Educational Testing Service, 1975.

BEST COPY

Lord, F.M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Mills , C.N., & Barr, J.E. A comparison of standard setting methods: Do the same judges establish the same standards with different methods? Paper presented at The Annual Meeting of the American Educational Research Association, Montreal, 1983.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Petersen, N.S., Marco, G.L., & Stewart, E.E. A test of the adequacy of linear score equating models. In P.W. Holland & D.B. Rubin (Eds.), Test Equating. New York, NY: Academic Press, 1982.

Potthoff, R.F. Some issues in test equating. In P.W. Holland & D.B. Rubin (Eds.), Test Equating. New York, NY: Academic Press, 1982.

Saunders, J.C., Ryan, J.P., & Huynh, H. A comparison of two approaches to setting passing scores based on the Nedelsky procedure. Applied Psychological Measurement, 1981, 5, 209-467.

Shepard, L.A. Setting standards. In M.A. Bunda & J.R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington, DC: National Council on Measurement in Education, 1979.

Shepard, L.A. Standard setting issues and methods. Applied Psychological Measurement, 1981, 4, 447-467.

Shimberg, B. Occupational licensing: A public perspective. Princeton, NJ: Educational Testing Service, 1982.

Shimberg, B., Esser, B., & Kruger, D. Occupational licensing: Practices and policies. Washington, DC: Public Affairs Press, 1972.

Skaggs, G., & Lissitz, R.W. Test equation: Relevant issues and a review of recent research. Paper presented at The Annual Meeting of the American Educational Research Association, New York, 1981.

Slinde, J.A., & Linn, R.L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Warm, T.A. A primer of item response theory (CG-941278). U.S. Coast Guard Institute: P.O. Substation 18, Oklahoma City, Oklahoma, December 1978.

Wright, B.D., & Stone, M.H. Best test design: Rasch measurement. Chicago, IL: Mesa Press, 1979.